Session 12

NONRESPONSE IN SURVEYS

EXPLORING NONRESPONSE IN U.S.FEDERAL SURVEYS

Maria Gonzalez, OMB; Dan Kasprzyk, NCES; and
Fritz Scheuren, IRS

Section 1: Introduction

This paper is intended to provide a broad summary of nonresponse rate trends in U.S. federal government surveys. We have built directly on the work of a Subcommittee on Survey Nonresponse, commissioned in 1991, by the Office of Management and Budget's Federal Committee on Statistical Methodology (FCSM). A particular debt of gratitude needs to be acknowledged for the role played by Bob Groves (Subcommittee Chair), Mick Couper and the other members of that Subcommittee for their input into what follows (see acknowledgements for a full list of the members).

Highlights of the Subcommittee's efforts have already appeared in the April AMSTAT NEWS (Gonzalez, Kasprzyk, and Scheuren,1994). A more extended treatment will be given in this paper. Still other papers based on the Subcommittee's work will appear in the Proceedings of the 1994 meetings of the ASA.

The present material is organized into four main sections, along with supporting figures, references, acknowledgements and an afterword. First, there is this short Introduction (Section 1); some background considerations come next. These considerations led to the establishment of the FCSM Nonresponse Subcommittee (Section 2).

In Section 3, an overview of the work of the Subcommittee is given, including the principal findings on nonresponse rate trends in federal surveys. Naturally, a discussion is given of limitations as well.

Finally, the recommendations of the Subcommittee are revisited in Section 4 and comments made on the future steps

that we, as federal government statisticians, should take -- both as individuals working on our own surveys and by acting collectively to improve practice as a whole.

Section 2: Some Background on Nonresponse in Federal Surveys

Like the poor, nonresponse in surveys may always be with us. In the days of "representative" samples drawn purposively, nonresponse was present but not visible. (Quota sampling, even today, makes measuring the extent of the actual nonresponse difficult -- maybe impossible). With the ascent of the random sampling paradigm (Bellhouse,1988), nonresponse became a problem that needed to be "solved."

In so far as U.S. Federal surveys are concerned, the turning point in government practice for the randomization paradigm came when Deming invited Neyman to lecture at the U.S.D.A. Graduate School in 1937. Morris Hansen, using Neyman's ideas and his own, and with many collaborators, did the rest (e.g., Hansen, Hurwitz, and Madow,1953).

It seems clear that Hansen and the other early pioneers understood quite well that randomization-based inference was directly challenged by nonresponse. Concerns about bias, for example, were evident from the beginning. In Cochran(1977) there is an example of an early treatment that simply widens the confidence intervals directly to account for the nonresponse bias. This conservative approach was consistent with the main focus of the random samplers of that era who were busy inventing ways to reduce nonresponse to the bare minimum. The U.S.Census Bureau in its Current Population Survey (e.g., Hanson,1978) still continues successfully in that tradition.

Hansen and his collaborators, in addition to a primary emphasis on "prevention," developed designs which called for the subsampling of nonrespondents (e.g., Hansen and

Hurwitz,1946). These were a natural extension of the basic randomization paradigm and called for more thorough fieldwork on a random subsample of nonrespondents. One of the results of this work was to introduce the idea of a weighted response rate. Such samples naturally also had their own nonresponse problems; so this approach too was seen from the beginning as only a partial one. Post-survey adjustment techniques to compensate for flaws in the randomization due to nonresponse were also attacked as well.

For those interested in more information, a special September 1975 issue of the Journal of the American Statistical Association is a recommended reference (Gonzalez, Ogus, Shapiro, and Tepping, 1975). This article provides a useful summary of federal goverment (largely Census Bureau) practices on the reporting of sampling and nonsampling errors, including nonresponse (see also Duncan and Shelton (1978) for still more on the history of sampling in U.S. Federal surveys).

While nonresponse in federal surveys has always been said to be an indicator of the quality of survey data, interest and concern has grown during the last two decades:

- The Panel on Incomplete Data, established by the Committee on National Statistics in 1977, produced three volumes focussing on incomplete data in sample surveys (Madow, Nisselson, Olkin, and Rubin, 1983).

- The Council of American Survey Organizations (CASRO) reviewed response rate definitions with the intent of trying to establish uniformity of definitions across surveys (CASRO, 1982).

- Steeh (1981) and Groves (1989) reviewed trends in the response rates in nongovernment surveys, indicating a decline in response rates over time.

- During the last ten years, the tight federal budget climate has prompted questions about the ability of federal statistical agencies to maintain high response rates with a constant budget.

Theoretical developments in the handling of nonresponse have grown enormously since the mid-1970's. Indeed, the problem has drawn the attention of some of the best statisticians now working on surveys. The National Academy Panel's report on Incomplete Data(1983) was a culmination of sorts. A review of nonresponse adjustment techniques was done by Kalton (1983). Even so, in the ten years since the Panel's report, there has been a lot more done and no end is in sight. The book on nonresponse by Little and Rubin (1986) and a separate book by Rubin (1987) on multiple imputation are perhaps the two most prominent examples of the important work that continues. The treatment of Sarndal et al (1992) and Lessler and Kalsbeek (1992) also are valuable for the way, among other things, they place nonresponse in context of total survey error.

Within this general environment of greater interest in nonresponse, the FCSM decided to sponsor an effort to learn what was known about nonresponse as a source of bias in federal survey estimates. Prominent factors in making this decision were --

- The lack of a systematic review of the topic since the 1983 Committee on National Statistics report.

- A growing perception among the members of the federal statistical community that nonresponse in federal surveys had been increasing over time.

In any event, in 1991 a Subcommittee of the FCSM was formed to study nonresponse in federal surveys. The initial charge of the Subcommittee was, simply stated, to "begin an effort to better understand unit nonresponse in surveys." The proposed approach was to conduct a broad-based review of the level of unit nonresponse rates, currently and over time, in federal surveys. The details of the Subcommittee's work are covered in the next Section.

Section 3: Work of FCSM Subcommittee on Survey Nonresponse

The Subcommittee was specifically charged with the mission to investigate for Federal surveys the levels of response rates, the measures used to compute these response rates, response trends from 1982-1991, perceived correlates of nonresponse, and other related information.

In carrying out its mission, the Subcommittee obtained information from 26 demographic and 21 establishment surveys. These surveys were not selected by probability methods, because no machine-readable listing of Federal surveys with sufficient auxiliary information for appropriate stratification was available. The 47 surveys were chosen, however, to include Federal surveys that differed on a number of key design parameters: those conducted on an ongoing or an intermittent basis, those conducted by Federal agencies, and those carried out by contractors under Federal auspices.

Because of the large differences in the design of surveys to collect establishment versus demographic data, separate questionnaires were constructed for each type and sent to respective survey sponsor or data collection agency. The intent of both questionnaires was to elicit information on a

variety of survey features that earlier literature has shown to affect nonresponse. In addition, information was sought on strategies for post-survey adjustment for nonresponse.

The Study itself incurred no unit nonresponse but did incur a small amount of item nonresponse in its data collection activities. Indeed, it was difficult to get the agencies to respond to the nonresponse questionnaire.

The findings of the Subcommittee span the range from the expected to the surprising. As in any research undertaking, of course, the conclusions drawn from an analysis of the questionnaires should be treated with caution. This point is particularly well taken here given the purposive nature of the sample, the small number of surveys included in the data collection, and the wide variety of design differences that characterize these surveys. Some highlights follow.

Trends in Nonresponse Rates. Despite the prior beliefs of many in the Federal survey community, there was little evidence of declining response rates over time for either the establishment or demographic surveys included in the Subcommittee's study:

- Establishment Surveys. To analyze the response rates of establishment surveys over time, it is more meaningful to limit the analysis to those surveys which reported response rates for several years. For this reason, the analysis of time trends in the response rates for establishment surveys cover only the nine surveys for which both weighted and unweighted data were available for at least six reporting periods between 1981-1991.

Figure 1 shows the average weighted response rate for the nine selected surveys. Figures 1-5 are based on work of the FCSM Subcommittee on Survey Nonresponse. As may be seen, the weighted response rate was only slightly decreasing over the period covered by the data. The average decrease was about 1/4 percent per year. Figure 1 also shows the mean unweighted response rate for the selected nine surveys from 1984-1990. The unweighted rate was slightly increasing, but stable over the period. The average increase was about 1/2 percent per year.

Figure 2 shows weighted response rates for the nine establishment surveys. Five of these weighted response rates are 90 percent or above. Two series have a weighted response rate between 70-90 percent and two series are around 50 percent. More about establishment trends in response will be said in Osmint, McMahon, and Martin.

- Demographic Surveys. Most demographic surveys used unweighted response rates rather than weighted rates for routine monitoring of the data collection process and so we have followed this convention here as well. The analysis of trends over time for demographic surveys was restricted to those surveys with at least 4 data points in the period 1982 to 1991. Only 8 of the 26 demographic surveys included in our data collection met this criterion.

The mean nonresponse rate by year was calculated for these eight surveys from the data provided, along with refusal rates and noncontact rates where available. Although the stimulus for the creation of the Subcommittee was the belief that response rates were

declining over time in demographic surveys, Figure 3 does not support that belief. The mean nonresponse rates for the surveys included in the sample are minimally lower in 1991 than in 1982. This figure shows that refusal rates, a major component of nonresponse, have remained about the same. More about demographic trends in response will be said in Johnson, Botman, and Basiotis.

For the Current Population Survey a longer time series of data is available. Figure 4 shows that the level of nonresponse has been stable for some years. Since the refusal rates seem to have increased, a possible implication is that more effort may have been made to reduce other nonresponse components -- so as to achieve relatively constant overall response rates.

An examination of response rates for the more-frequently fielded demographic surveys reveals large variations across surveys. This variation can be partially understood by separating the studies into two groups (see Figures 5). One group has response rates in the 95 percent range, while a second cluster lies about 10 percentage points lower. The studies in the 95 percent range consist of ongoing studies conducted by the same interviewer corps. The studies in the lower group tend to be less frequently conducted. Neither group exhibits strong trends over time.

In summary, despite the prior beliefs of many in the survey community, there was little evidence of declining response rates over time among either the establishment or demographic surveys included in the study. This could be due to a greater effort in data collection but technological and other survey context changes make this hard to verify. One

final point about these results may be worth making again: There were only a limited set of surveys on which time trends can be measured--just nine establishment surveys and eight demographic surveys.

Other Findings.-- There are other findings from the Subcommittee's work; but only three are highlighted here. These involved issues in the definition of nonresponse, response rate documentation, and post-survey adjustment methods:

- Definitions for Nonresponse. Despite the study's focus on nonresponse rates and despite having contacts in the agencies, major difficulties arose in obtaining consistent information. Just as was found in an early (albeit more general) study,..."rates have different names and different definitions in different places and times." (Bailar and Lanphier, 1978) This issue led, in part, to one of the study's major recommendations (see figure 6, Subcommittee Recommendation 3).

- Response Rate Documentation. Reporting practices for documenting response rate components varied widely across the surveys in the study. Common practice in establishment surveys is in contrast to common practice in demographic surveys. Sponsors of demographic surveys not only were more likely to maintain records regarding a wider variety of nonresponse components but also tended to maintain more historical information. For example, all of the demographic surveys in our data collection included some information about response/nonresponse components. In contrast, for the establishment surveys analyzed, 10 out of 21 did not track any nonresponse components.

- Post-survey Nonresponse Adjustment. Respondents were asked about a number of post-survey adjustment techniques designed to reduce the effects of nonresponse: post-stratification (e.g., simple ratio or raking ratio adjustment), regression modelling of the propensity to respond, and imputation. All surveys in the Subcommittee study used some degree of post-survey nonresponse adjustment. Some of the approaches were very traditional, while others reflected more recent research on estimation strategies.

In the remaining section of this paper we cover the Subcommittee's recommendations and a few ideas on future steps.

Section 4: Some Next Steps for Practice

The Subcommittee made four recommendations that are given in detail in figure 6. Stated briefly the subcommittee recommended:

- Survey practitioners should compute nonresponse rates in a uniform fashion over time.

- In repeated surveys, response rate components should be monitored in conjunction with cost and design changes.

- Agencies that sponsor surveys should publish how they compute response rate and their components in survey reports and their relevance to the quality of the survey results discussed.

- Ongoing research should be conducted on nonresponse adjustment variables, costs and benefits of converting refusals, and similar nonresponse management concerns.

All of these recommendations seem rather obvious. They address some very basic survey management and reporting requirements; furthermore, the suggestions are close in spirit and substance to those made by other groups over the last two decades.

It is true that every survey program examined by the Subcommittee calculated nonresponse rates in some fashion and had some auxiliary information about aspects of nonresponse. It is also true that most survey programs did not have **readily** available what the Subcommittee viewed as "basic" data on nonresponse; nor did repeated surveys have a time series easily available of nonresponse rates and nonresponse components.

What can we expect for the future based on the results of this small exploratory study? Some conjectures follow:

- First, it is unrealistic to assume that the recommendations by yet another subcommittee will be adopted uniformly by the agencies of the Federal statistical system.
- Second, unless mandated, individual survey program managers are likely to remain individualistic and independent with respect to their acceptance and adoption of recommendations concerning their surveys.
- Third, it is important to recognize the diversity of the management of individual survey programs and build on each survey programs' strengths. In other words, these recommendations should be no more than guidelines in any case.
- Fourth, the survey data collection manager and the agency that sponsors the survey need to work together as a team with the interests of the ultimate customer paramount -- recognizing that an information system producing data on nonresponse and its components is mutually beneficial.

One need not assume the points made above are necessarily pessimistic. The underlying theme is the development of a fully professional partnership among data collection managers, agencies that sponsor survey programs, and ultimate customers. Mutual respect and understanding for each other's requirements (given budget constraints) is essential for improving the reporting of nonresponse and nonresponse components. The current theme of "reinventing government" speaks well to the prospect of improving these professional relationships through its team building and customer orientation emphases.

Finally, we can expect incremental improvements in the issues discussed here through the continuing work of the Federal Committee on Statistical Methodology (Gonzalez, 1994) and the National Science Foundation initiated "Program in Survey Methodology" offered by the consortium of the University of Maryland, University of Michigan, and Westat. Both programs are dedicated to the improvement of the quality of Federal survey data. Through these efforts and the individuals involved in Federal data collection programs, progress will be made.

Most of this paper looks inward at the federal statistical system. Obviously, much can be learned by examining private sector experiences and through international comparisons. The companion paper at this session by David Binder and his colleagues from Statistics Canada is an example of what we have in mind. Clearly, too, the Statistics Canada approach to nonresponse rates is worthy of further study by those interested in this area (see Statistics Canada, 1993 and Hidiroglou, Drew, and Gray, 1993)

In the spirit of "reinvention," a systematic benchmarking approach is needed. Some important beginning efforts that bear mention in this regard include the papers by Lyberg and Dean (1992) and Christianson and Tortora (1993).

Acknowledgements and Afterwords

This paper is based on the work of the Subcommittee on Survey Nonresponse whose members included: Robert M. Groves, Chair, BOC, Susan Ahmed, NCES, J. Donald Allen, NASS, David Belli, BEA, Peter Basiotis, HNIS, Steve Botman, NCHS, Eileen Collins, NSF, Mick Couper, BOC, Patricia Guenther, HNIS, Paul Hsen, BLS, Ayah Johnson, AHCPR, Arthur Kennickell, FRB, Paul McMahon, IRS, Jeffrey Osmint, BoM, Antoinette Martin, EIA, Pamela Powell-Hill, BOC, Maria Reed, BOC, Carolyn Shettle, NSF.

One final point may be made about the exploratory nonresponse Study focussed on in this paper. We would like to engage in a dialogue about this study and its recommendations with those who conduct both Federal and non-Federal surveys. Please contact Maria E. Gonzalez (OMB, Statistical Policy Office, NEOB, Room 10201, Washington, D.C. 20503) with your comments or experiences in this field.

References

Bailar, B.A. and Lanphier, C.M. (1978). Development of Survey Methods to Assess Survey Practices. Washington: American Statistical Association.

Bellhouse, David (1988). "A Brief History of Random Sampling Methods," Handbook of Statistics, Vol 6, North-Holland, Amsterdam, pp. 1-14.

Christianson, Anders and Robert D. Tortora (In Press). "Issues in Surveying Establishments: An International Survey," in BG Cox et al (Editors), Survey Methods for Businesses, Farms, and Institutions, John Wiley & Sons. Presented at the International Conference on Establishment Surveys, 1993 Buffalo, NY.

Council of American Survey Organizations (CASRO) (1982). On the Definitions of Response Rates, Port Jefferson, New York.

Duncan, J.W. and Shelton, W.C. (1978). Revolution in United States Government Statistics 1926-1976, U.S. Department of Commerce, Office of Federal Statistical Policy and Standards.

Gonzalez, M.E. (1994). "Improving Data Awareness in the United States Statistical Agencies," The American Statistician. American Statistical Association. Vol. 48, No. 1.

Gonzalez, M.E., Kasprzyk, D., and Scheuren, F. (1994). Nonresponse in Federal Surveys, AMSTAT News, April.

Gonzalez, M.E., Ogus, J.L., Shapiro, G., and Tepping, B.J. (1975). Standards for the Discussion and Presentation of Errors in Census and Survey Data, Journal of the American Statistical Association, 70, 351, Part II.

Groves, Robert M. (1989). Survey Errors and Survey Costs, John Wiley and Sons, New York.

Hansen, M., Hurwitz, W., and Madow, W. (1953). Sample Survey Methods and Theory (2 Volumes), John Wiley: New York.

Hansen, M. and Hurwitz, W. (1943). On the Theory of sampling from finite populations. Ann. Math. Statist. 14, 333-362.

Hanson, R. (1978). The Current Population Survey--Design and Methodology, Technical Paper 40, U.S. Bureau of the Census.

Hidiroglou, M.A., Drew, J.D., and G.B. Gray (1993). "A Framework for Measuring and Reducing Nonresponse in Surveys," Survey Methodology, Vol. 19, No. 1, pp. 81-94.

Johnson, Ayah, Steve Botman, and Peter Basiotis (forthcoming). "Nonresponse in Federal Demographic Surveys: 1981-1991." To be presented at the 1994 annual meetings of the American Statistical Association, Section on Survey Research Methods.

Kalton, G. (1983). Compensating for Missing Survey Data, Research Report Series, Ann Arbor, Michigan: Institute for Survey Research.

Lessler, J. and Kalsbeek, W. (1992). Nonsampling Errors in Surveys, John Wiley & Sons: New York.

Little, R. and Rubin, D. (1986). Statistical Analysis and Missing Data, John Wiley & Sons: New York.

Lyberg, L. and Dean P. (1992). "Methods for Reducing Nonresponse Rates: A Review," presented at the annual meeting of the American Association for Public Opinion Research, St. Petersburg, FL.

Madow, W.G., Nisselson, H., Olkin I., and Rubin D.B. (Editors) (1983). Incomplete Data in Sample Surveys, 3 Volumes, Academic Press, NY.

Osmint, Jeffrey, Paul B. McMahon, and Antoinette Ware Martin (forthcoming). "Response in Federally Sponsored Establishment Surveys." To be presented at the 1994 annual meetings of the American Statistical Association, Section on Survey Research Methods.

Rubin, D. (1987). Multiple Imputation for Nonresponse in Surveys, John Wiley & Sons: New York.

Sarndal, C.E., Swenssen, B., and Wretman, J. (1992). Model Assisted Survey Sampling, Springer-Verlag: New York.

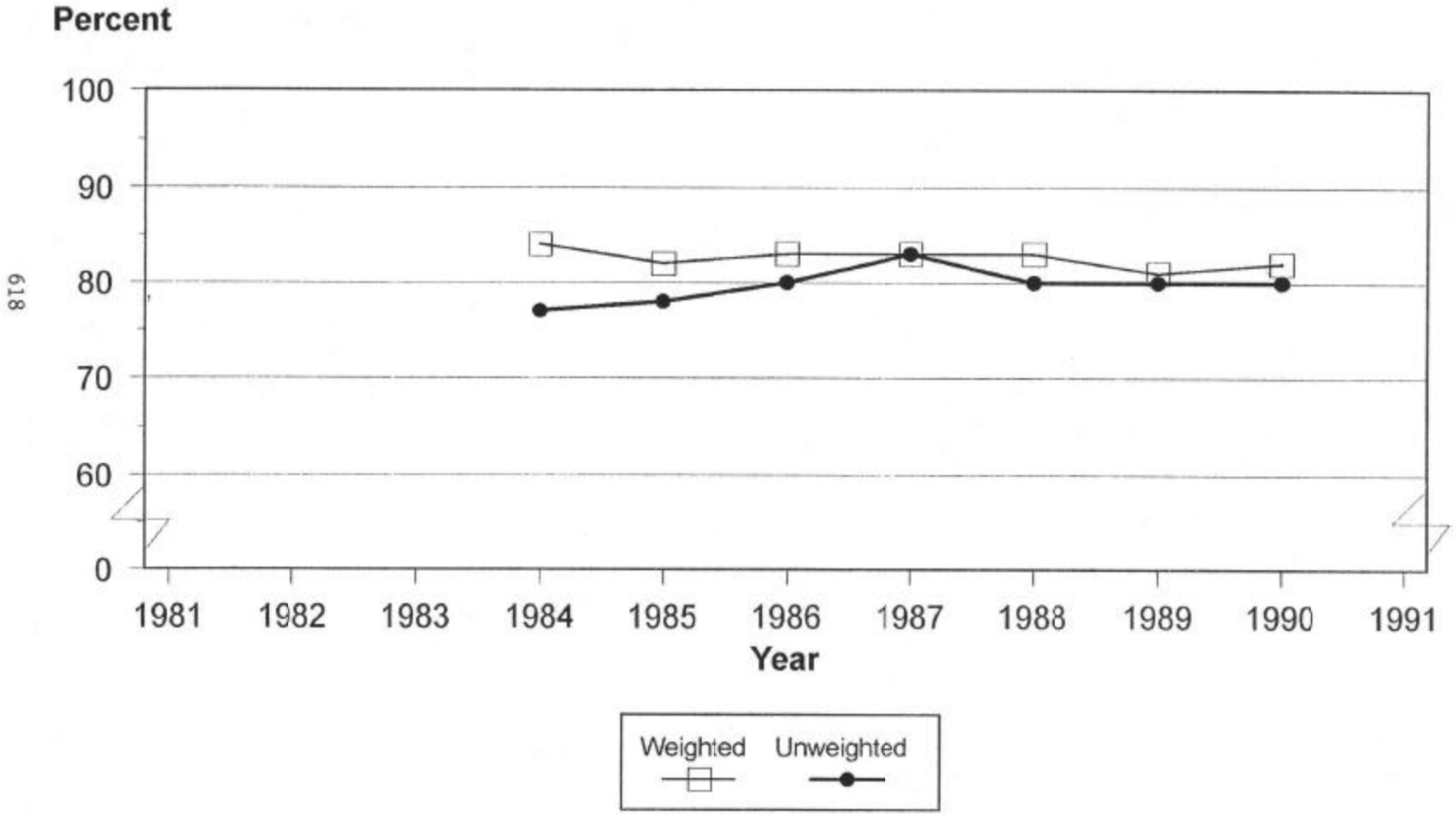
Shettle, Carolyn F., Patricia M. Guenther, Daniel Kasprzyk, and Maria Elena Gonzalez (forthcoming). "Investigating Nonresponse in Federal Surveys." To be presented at the 1994 annual meetings of the American Statistical Association, Section on Survey Research Methods.

Statistics Canada (1993). Standards and Guidelines for Reporting of Nonresponse Rates, Ottawa, Canada.

Steeh, C. (1981). "Trends in Nonresponse Rates," Public Opinion Quarterly, Vol. 45, pp.40-57.

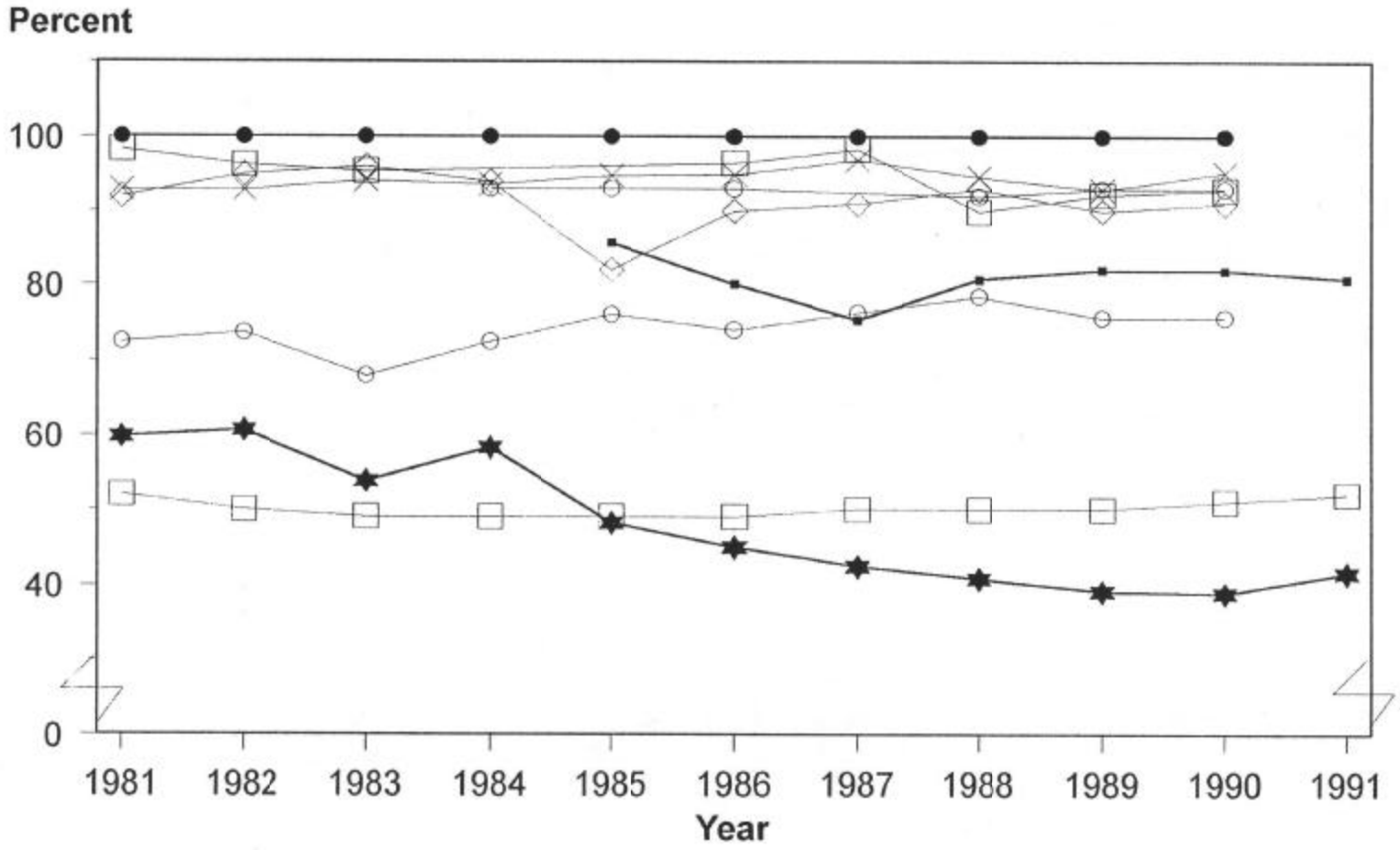
Figure 1.

Mean Response Rates, Economic Surveys



SOURCE: Federal Committee on Statistical Methodology, Subcommittee on Survey Nonresponse.

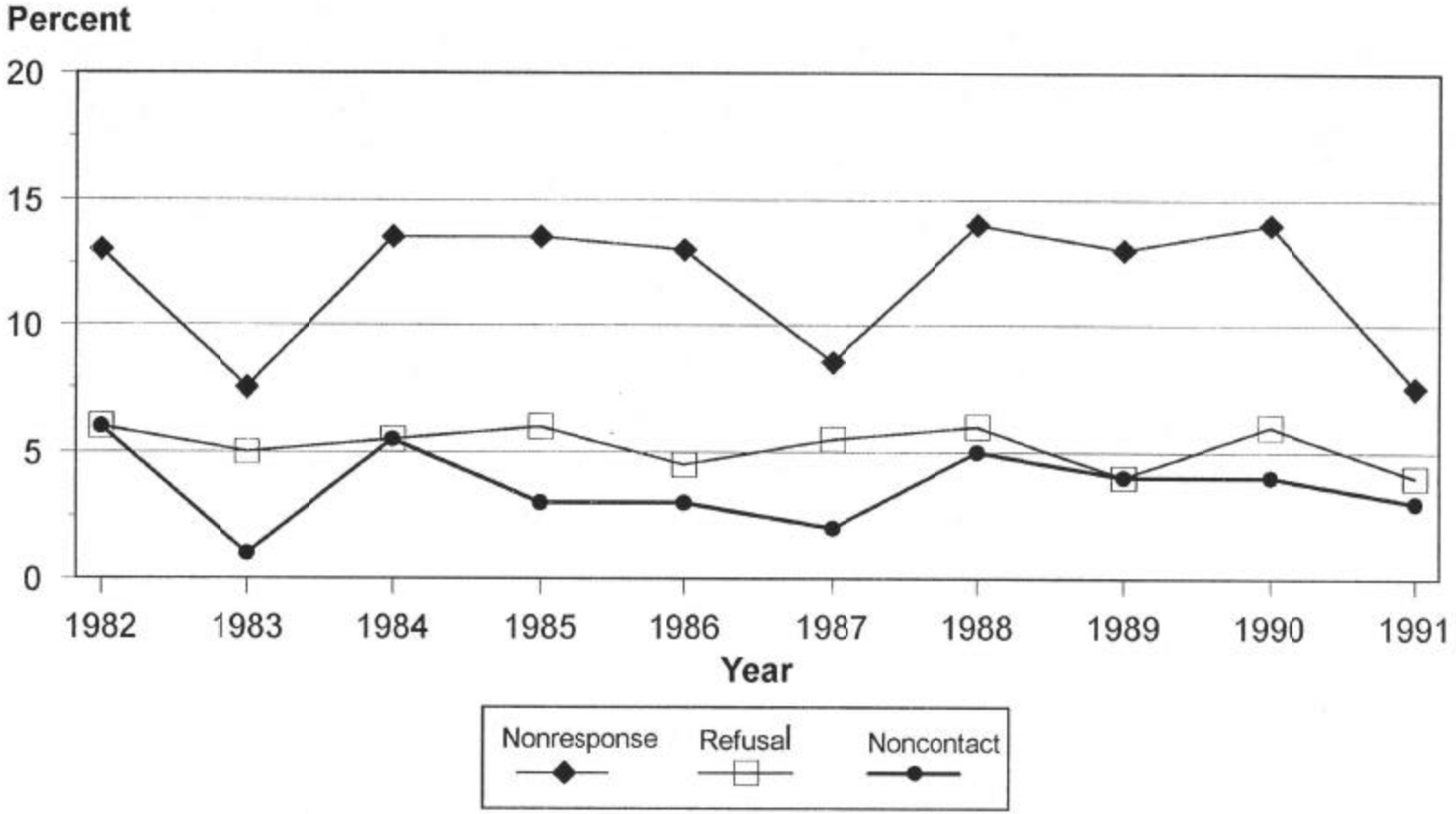
Figure 2.
Weighted Response Rates, Economic Surveys



SOURCE: Federal Committee on Statistical Methodology, Subcommittee on Survey Nonresponse.

Figure 3.

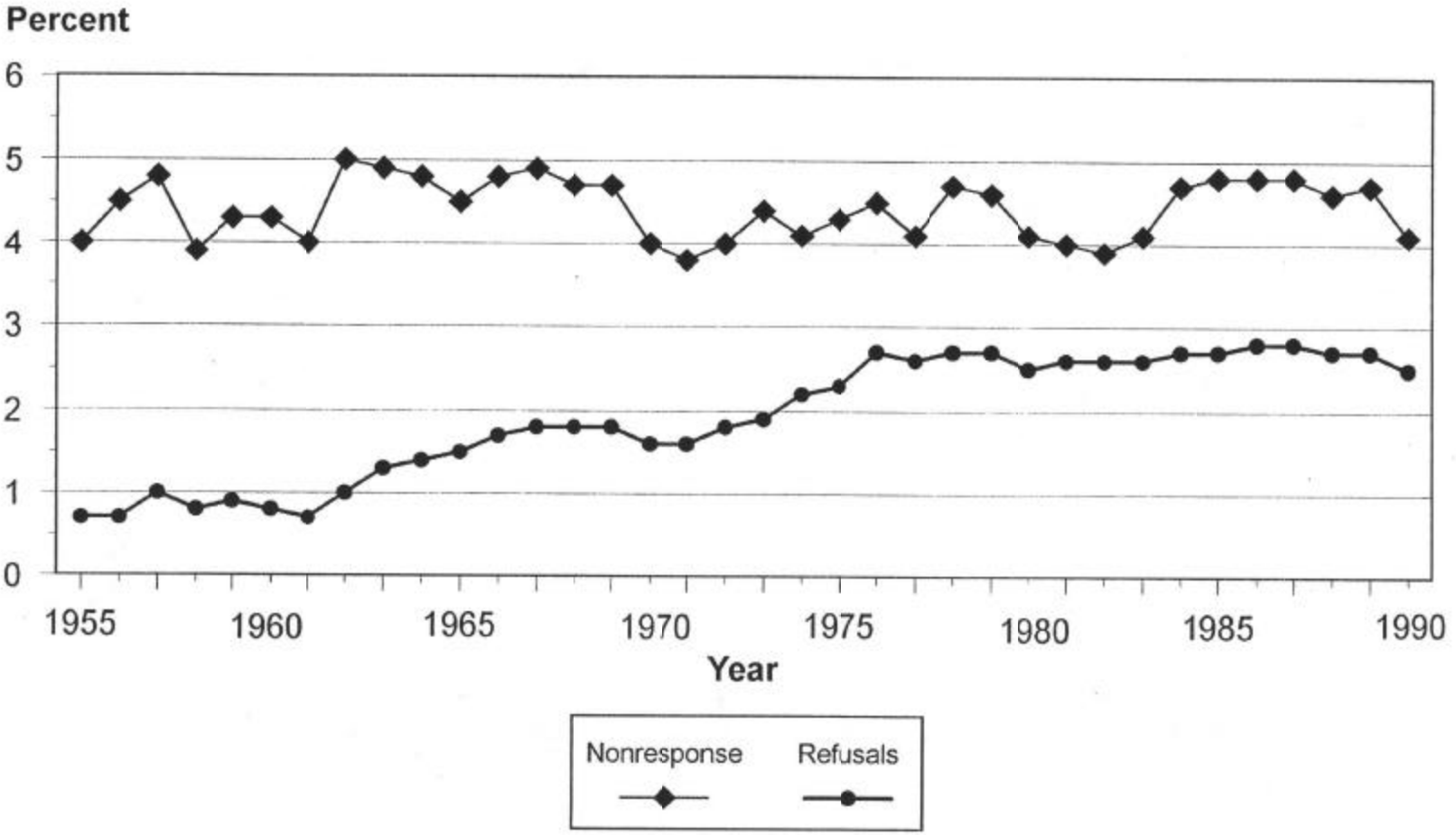
Mean Unweighted Nonresponse Rates, Demographic Surveys



SOURCE: Federal Committee on Statistical Methodology, Subcommittee on Survey Nonresponse.

Figure 4.

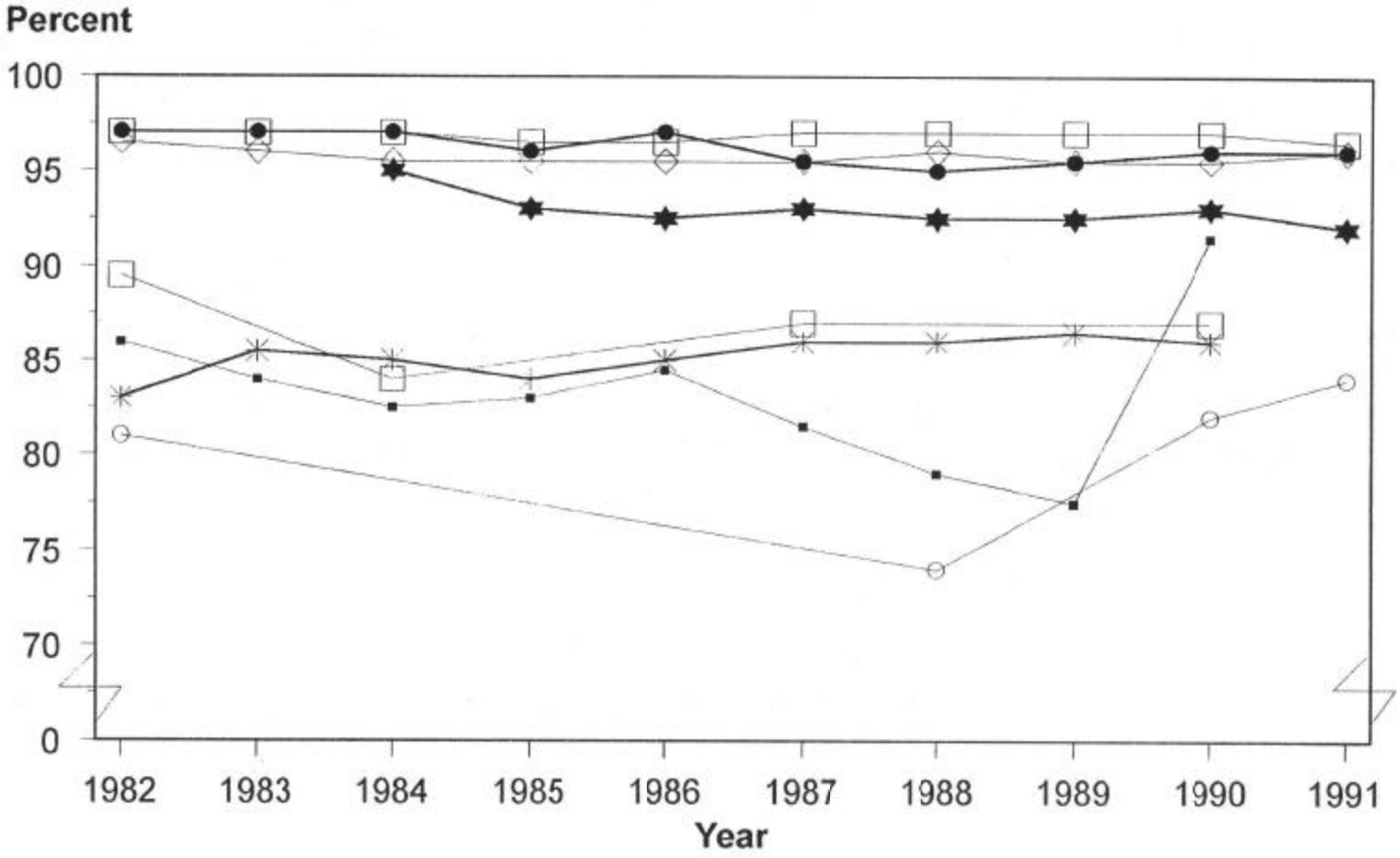
Unweighted Nonresponse Rates, Current Population Survey



SOURCE: Federal Committee on Statistical Methodology, Subcommittee on Survey Nonresponse.

Figure 5.

Unweighted Response Rates, Demographic Surveys



SOURCE: Federal Committee on Statistical Methodology, Subcommittee on Survey Nonresponse.

Figure 6.--Summary of FCSM Recommendations on Survey Nonresponse

Recommendation 1. **Survey staffs should compute response rates in a uniform fashion over time and document response rate components on each edition of a survey.**

The subcommittee chose not to recommend that every survey use the same response rate computations. Other groups have recommended such uniformity (see CASRO, 1982). In the Subcommittee's view, every definition of response rate components offers some useful information. Some response rate definitions inform the designers about the rate of success of measurement of the average sample unit; others focus on different causes of nonresponse. One can distinguish between measures useful as management tools and measures that should be reported to data users so that they can assess the quality of the survey data.

Recommendation 2. **Survey staffs for repeated surveys should monitor response rate components (e.g., refusals, not-at-homes, out-of-scopes, address not locatable, postmaster returns, etc.) over time, in conjunction with routine documentation of cost and design changes.**

The Subcommittee believes that response rate components are useful tools to monitor changes in the quality of survey statistics. Response rates should be easily accessible and timely. By themselves, they are not error measures; however, for repeated surveys, changes in response rate components may signal the need for supplementary study of nonresponse error properties. Such changes can alert the survey designers to changes in the "survey-taking climate" that affect completion of measurement, point to changes in the administrative controls over response rates that may need adjustment, and help measure the effects of any design changes made.

For ongoing surveys, graphs of time series of response rate components, juxtaposed with costs for each collection cycle, and indicators of design changes introduced in that cycle, can be valuable management tools. Survey managers need better tools to diagnose the causes of cost changes in data collection activities. Falling response rates, especially those associated with cases requiring much effort prior to the ultimate nonresponse, magnify cost pressures on surveys. The subcommittee's study did not collect data on survey costs, because comparable cost information across surveys was not believed to be available.

Recommendation 3. **Agencies that sponsor surveys should be empowered to report the response rates of their surveys. The sponsoring agency should explain how response rates are computed for each survey it sponsors. Response rates for any one survey should be reported using the same measures over time, so that**

users may compare the response rates. Response rate components should also be published in survey reports.

An agency that sponsors surveys should compute and explain in its survey publications the response rates for each of the surveys it sponsors. Surveys sponsored over time should report the same measure of response for all data collection periods so that users can compare these measures over time. The actual method used to compute response rates should be described in all publications issued.

The results of recommendations 1 and 2 should be shared routinely with the users of survey data, along with discussions of the relevance of response rates to evaluating the quality of the survey data. An analysis of the characteristics of the nonrespondents should be implemented routinely as part of each cycle of data collection.

Recommendation 4. **Some research on nonresponse can have real payoffs. It should be encouraged by survey administrators as a way to improve the effectiveness of data collection operations.** The Subcommittee believes that areas of research most likely to yield payoffs include:

- Studies of the relative costs of final efforts to raise response rates, through persuasion, repeated callbacks, and other measures. When these costs are compared to number of cases added to the respondent pool, the relative cost per case can be computed. Studies of the effects of these final cases can be made in an effort to assess the cost effectiveness in terms of mean square error of the final efforts.

- Studies of the measurement error properties of information provided by the reluctant respondent cases, relative to the nonresponse bias in statistics that would omit them from computations. This would address a key question in survey design: When data collectors exert great effort to persuade the reluctant to respond, is one type of error, nonresponse, merely exchanged for another type, measurement error? Perhaps, those persuaded to respond may exert less effort at providing accurate data?

- Studies on what variables should be collected to improve post-survey adjustment for unit nonresponse (see Madow et al, 1983: Recommendation 10(2)). When observable or inferred characteristics of nonrespondent units are related to the survey variables and to the likelihood of participation, then collecting and using these variables in post-survey adjustment models might be a cost effective method of reducing overall mean square errors.

MODEL-BASED REWEIGHTING FOR NONRESPONSE ADJUSTMENT

David A. Binder, Sylvie Michaud and Claude Poirier
Statistics Canada, Ottawa, Ontario, Canada K1A 0T6

ABSTRACT

Nonresponse in surveys is inevitable. Much has appeared in the literature on methods of compensating for this source of nonsampling error. There is a growing interest in attempting to understand the causes of nonresponse and studying the differences in characteristics between respondents and nonrespondents. In this paper, we briefly review some related literature, discuss modelling approaches for adjusting for nonresponse and present the research findings for two surveys conducted at Statistics Canada. In both the Survey on Labour Income Dynamics and the Farm Financial Survey, we examine differences in characteristics between respondents and nonrespondents and the suitability of adopting a modelling approach for compensating for nonresponse.

KEY WORDS: Generalized regression estimators; Logistic regression; Response propensity.

1. INTRODUCTION

In virtually every survey, no matter how carefully it is designed, we must accept the fact that some data will be missing. Other than data that is missing by design, such as data from nonsampled units, data can be missing for many reasons; for example, non-contact with the respondent, refusals, late reporting, collection and processing errors, data deletion due to edit failure, undercoverage, etc.

Some measures must be taken to deal with such nonresponse. Over the years, a host of techniques has been developed. The actual choice of technique should depend on a number of factors. These include the method of estimation to be used, the amount of information about the nonrespondents that is available, the extent of other sources of error such as sampling error and response error, the relative importance of the variables to be estimated, the resources available for exploring the problem, the nature of the analyses to be performed and the statistical inferences to be made from the survey, etc.

However, even with all of these criteria, there must necessarily be some subjective judgments on the nature of the nonresponse. As we shall see, many of the methods for coping with nonresponse make use of models, either explicitly or implicitly. Even the most ardent advocates of the pure design-based school will resort to some model assumptions when it comes to adjusting for nonresponse. This presents a new set of problems associated with the statistical inferences, since the randomization distributions on which the inferences are based are no longer purely design-based, unless the nonresponse mechanism can be considered to be part of that design.

In this paper, we shall focus on the implications of the estimation method to be used and the amount of information about the nonrespondents that is available. It will be assumed that the

prime focus of the survey is to obtain estimates of descriptive statistics, such as means, totals, differences and ratios. Often nonresponse is broadly categorized into *unit-level* nonresponse and *item-level* nonresponse. This categorization is often extended, in the case of longitudinal or follow-up surveys, to *wave-level* nonresponse, where wave nonresponse is usually unit nonresponse on a particular survey occasion. In fact, it is for the case of wave nonresponse where we have the richest source of data for the nonrespondents who reported in previous waves.

Unit nonresponse is usually defined as cases where only the frame information is available for the respondents. In practice, this definition is extended to other cases where there is insufficient usable data from the respondents. The usual method for dealing with unit nonresponse is to use an "appropriate" weighting procedure to compensate for the nonresponse. (We define weighting procedures here broadly to include weight adjustments implied by regression, ratio or similar estimation techniques using auxiliary data.)

On the other hand, item nonresponse is handled either through imputation at the item level, or by ignoring the usable information and treating the respondent like unit-level nonrespondents. For wave-level nonresponse to longitudinal surveys, either reweighting or item imputation may be suitable. In this paper, we focus on the methods that use weighting techniques.

In Section 2, we discuss the basic theory underlying many of the adjustment methods and give a brief literature review. In Sections 3 and 4, we give examples of two surveys at Statistics Canada where some of these models have been studied recently. We summarize our findings in Section 5.

2. SOME GENERALITIES

2.1 Estimation

In general, we are interested in means, totals, ratios, etc. of survey variables. We denote the value of the i-th survey variable for the k-th respondent as y_{ik}. In cases where the occasion, t, is relevant, we can use y_{ikt} instead. A sample is selected according to some well-defined sampling plan. The sampling plan is usually based on frame information such as geography, and other classification and size variables. We use s to refer to the selected sample. Unfortunately, in practice, after the k-th respondent is selected, a number of things can go wrong in the process of obtaining and recording the y-values. Some of these are in the general category of response errors, where we obtain data, but they are not the y-values we were seeking. In this paper, we ignore these types of errors, except to point out that if these errors lead to large biases, the resources for nonresponse concerns may need to be trimmed in order to address the larger problem. The problem that we are addressing here is the case where the y-values are unobtainable. We denote by $s' \subseteq s$ the set of units for which we obtain usable y-values. (The subscript t is implied, where appropriate, for longitudinal surveys.)

First, we describe the estimators in the case of no nonresponse. Associated with each sampled unit, k, we have a survey weight given by

$$w_k(s) = g_k(s)\,\pi_k^{-1},$$

where π_k is $\Pr(k \in s)$, the usual first-order inclusion probability, and $g_k(s)$ is a weight adjustment that makes use of auxiliary frame data, such as poststratification, regression and ratio adjustments, etc.; see, for example, Särndal, Swensson, and Wretman (1992). We assume that the estimator of a total for a y-variable on the t-th occasion is given by

$$\hat{Y}_{it} = \sum_{k \in s_t} w_k(s_t)\, y_{ikt}. \tag{2.1}$$

Note that this estimator could be made more general, if necessary, to allow for composite estimators and multiphase samples which can depend on y-values that are observed on other occasions, but we do not introduce this complexity here. Sufficient conditions for (2.1) to be asymptotically design consistent are:

1) the probability distribution of s depends only on the auxiliary data but not directly on the y-values for the current occasion, (2.2.1)
2) the limiting expectation of $g_k(s)$ is unity, (2.2.2)
3) the variance of $\hat{Y}$ is asymptotically zero. (2.2.3)

We now consider the implications of nonresponse. Formally we assume that, given the sample, s, the set of responding units, s', follow a probability distribution $p(s'|s)$. This is completely general, allowing for correlated response patterns. It also allows for the classical case, where it is assumed that the response behaviour is nonrandom and is an inherent attribute of the selected respondents, just like the survey variables. We now consider methods of nonresponse adjustment which we refer to as *generalized reweighting methods*. Associated with each responding unit, k, we have an adjusted weight given by

$$w_k'(s', s) = g_k'(s', s)\, w_k(s),$$

where $g_k'(s', s)$ is a weight adjustment that makes use of auxiliary frame data, as well as other information that may be available for the nonresponding units. This allows the weight adjustment to depend on survey values that were observed on previous occasions from a longitudinal survey. We assume that the estimator of a total for a y-variable on the t-th occasion is given by

$$\hat{Y}_{it}^{(GR)} = \sum_{k \in s_t'} w_k'(s_t', s_t)\, y_{ikt}. \tag{2.3}$$

We let $\rho_k(s)$ be $\Pr(k \in s' | s)$. In addition to (2.2.1) to (2.2.3) above, sufficient conditions for (2.3) to be asymptotically consistent with respect to the original design and the response probabilities are:

1) the probability distribution of s' given s depends only on the auxiliary data and the survey data from previous occasions, but not directly on the y-values for the current occasion, (2.4.1)

2) the limiting expectation of $g_k'(s', s)$ is $\{E[\rho_k(s)]\}^{-1}$, (2.4.2)

3) the variance of $\hat{Y}^{(GR)}$ is asymptotically zero. (2.4.3)

If (2.4.2) is violated, then the expectation of $\hat{Y}^{(GR)}$ is

$$\sum E[g_k'(s', s)]E[\rho_k(s)]y_k. \qquad (2.4.4)$$

The form of this bias is important, because if one were to impose model assumptions on the y-variables, it is possible that the model-bias becomes small. However, for those who wish to make the fewest model assumptions, it is clear that one should restrict attention to adjustment methods which yield condition (2.4.2) as closely as possible. This implies that the weight adjustment should reflect the propensity to respond as nearly as possible. Of course, the probability mechanism generating these response probabilities are generally unknown, so the weight adjustment must necessarily be model-based.

Another important feature of (2.4.2) is that if there are some "hard-core" nonrespondents -- that is, units where $\rho_k=0$ -- there would be no consistent estimates.

2.2 Examples from the Literature

The most basic form of reweighting for nonresponse that may lead to acceptable results is to simply use $g_k(s')$ instead of $g_k(s)$ in (2.1). This implies that

$$g_k'(s', s) = g_k(s')/g_k(s).$$

This was suggested by Bethlehem (1988) for the case of the generalized regression estimators. In this case we have that the bias of the estimator is $X\beta^* - Y$, where β^* is the expected value of the estimated π-weighted regression coefficient with no nonresponse adjustment. We see then that even though this estimator is generally biased, if the regression model is reasonable, the bias can be small.

Oh and Scheuren (1983) discussed weighting class adjustment methods, which is a poststratified estimator using weighting classes as poststrata. We see that this is consistent under the assumption that the response propensities are equal within weighting classes. In practice, this technique is in widespread use; see, for example Chapman, Bailey, and Kasprzyk (1986). It was

extended to generalized regression estimators by Särndal and Swensson (1987).

One of the difficulties with weighting class adjustment methods is that there may be too many weighting classes to control. Binder and Théberge (1988) showed that with a multiplicative model for response propensities, raking ratio estimators will yield unbiased estimates. This is consistent with (2.4.2). More complex weighting schemes are proposed by Alexander (1987) and Deville, Särndal, and Sautory (1993). These could be justified under various model assumptions for the response propensities.

Many authors have proposed the use of logistic regression models to explain the nonresponse mechanism. This is a commonly used model for binary dependent variables. Examples of this can be found in Ekholm and Laaksonen (1991), Folsom (1991), and Lepkowski, Graham, and Kasprzyk (1989). In the latter paper, the logistic regression model is compared to weighting class adjustment methods, where the weighting classes are determined through some data analytic searching methods.

In Iannacchione, Milne, and Folsom (1991), after weights are included to reflect the estimated propensity to respond, the weights are fine-tuned so that certain estimates correspond to the estimate obtainable with the nonrespondents included. This is possible for wave nonresponse where certain estimates can be made for a previous wave using either the previous wave respondents or the current wave respondents. This technique should generally improve the estimates. The differences in the estimates can also be used as an diagnostic tool for the model.

Judkins and Lo (1993) and Eltinge and Yansaneh (1993) used logistic regression to model the nonresponse propensities, but then created weighting classes based on the fitted values and used weighting class adjustment methods to reweight. One of the drawbacks of the weighting class adjustment methods is that the appropriate weighting classes are not always obvious, so that such data modelling is used to help define the classes. It is expected that this method should yield results that are similar to the weights based on the logistic regression. However, if the logistic model is correct, the method will tend to introduce a small bias since (2.4.2) will be violated. In practice, though, the logistic regression model is only an approximation to the true probability mechanism.

As we can see, reweighting methods have a strong base in the literature. The theory we have given in Section 2.1 indicates that the validity of these methods are model-based. Therefore it can be important to study the characteristics of the nonrespondents to develop the most suitable model. In Sections 3 and 4, we perform such studies on each of two surveys. We see that the models help our understanding of the factors that contribute to nonresponse.

An important side benefit of such studies is to help the survey manager pinpoint areas for improvement in the data collection phase.

3. SURVEYS OF LABOUR AND INCOME DYNAMICS AND LABOUR MARKET ACTIVITY

3.1 Introduction

Statistics Canada launched a major panel survey of households in 1994 called the Survey of Labour and Income Dynamics (SLID). The survey follows individuals and families for six years, collecting information on their labour market experiences, income and family circumstances. Its origins are in several surveys, including the Labour Market Activity Survey (LMAS). The LMAS served both as a longitudinal and as a cross-sectional survey. Two panels have been conducted to date, a two-year panel during 1986-1987 and a three-year panel during 1988-1990. For each longitudinal panel, respondents who participated in the first wave were interviewed and traced. All persons living with them in the following waves were also interviewed but not traced. Different studies are currently being conducted on nonresponse to the LMAS in hopes of finding approaches that will minimize the impact of nonresponse on the SLID data. Here we discuss our study on model-based reweighting.

Similarly to its predecessor (LMAS), the longitudinal sample for SLID is selected from the sample of dwellings that participated in the Labour Force Survey (LFS) in January 1993. The LFS has a response rate of 95%. Out of those respondents close to 90% agreed to participate in SLID. This sub-sample of respondents, comprising 15,000 households, is defined as the longitudinal sample, representative of the Canadian population as of January 1993. The longitudinal sample will be interviewed for six years, with two interviews carried out each year. Note that a sub-sample of LFS respondents who had refused to participate to SLID has been selected for evaluation purposes. If they respond in subsequent years, we may be able to determine how different they are from the rest of the sample. Preliminary analysis could not find systematic differences in the LFS characteristics between the nonrespondents and the respondents. More studies will be done by linking the full sample to administrative files to be able to evaluate if there are differences in terms of income characteristics.

Attritional nonresponse will be compensated with a weighting adjustment. Imputation will be used to compensate for some nonresponse; for example, nonresponse that is non-attritional. The weighting will include the following steps:

i) calculation of the initial weight based on the sample design,
ii) nonresponse adjustment,
iii) post-stratification by province, age groups, and sex to the 1993 population estimates.

The longitudinal panel of LMAS has been used as the research vehicle for the nonresponse modelling and weighting adjustments.

3.2 LMAS Survey Design and Nonresponse

For the first interview of the panel, LMAS is conducted as a supplement to the January Labour Force Survey (LFS). All eligible respondents from the LFS are included in the LMAS sample. In the

subsequent waves, for the longitudinal component of LMAS, all respondents to the first wave are interviewed in January of the following year(s). People are traced if they have moved.

LFS uses a multiple stage sample design. A stratum is defined based on geographic variables. At least two distinct PSU's (primary sampling units) are selected within each stratum. LFS initial weights go through a series of adjustment factors at the stratum level to produce a sub-weight. This sub-weight is then adjusted to population estimates by province/age-group/sex groups, plus an adjustment by Economic Region and Census Metropolitan Area, to produce a final weight. More details may be found in Singh, Drew, Gambino, and Mayda (1990).

For the LMAS longitudinal sample, nonresponse adjustment is done at the stratum- component level, corresponding to a PSU or a group of PSU's, as defined for the LFS. A poststratification is then done to adjust the nonresponse adjusted weights to population estimates at the province/age-group/sex level.

When the LMAS file was evaluated, it was found that nonresponse was quite different among certain groups:

- movers, including people that could not be traced, had a nonresponse rate of close to 20% while nonresponse for non-movers was about 2%. This was by far the characteristic that presented the most differences,
- based on characteristics from Wave 1, persons that were employed in Wave 1 had higher response rates after three years than those who were unemployed in Wave 1,
- similarly, persons that were married in Wave 1 had higher response rates in Year 3, compared to those who were single in Year 1,
- persons who lived in non-urban areas in Year 1 had higher response rates after three years.

The different characteristics between respondents and nonrespondents suggested that nonresponse adjustments should be done at some level different than stratum-component. Logistic regression was used to model the nonresponse behaviour. The multiple logistic response function is

$$\text{logit}(p) = \log[p/(1-p)] = \boldsymbol{\beta}'\boldsymbol{x},$$

where p is the probability of response to the 1987 survey for a 1986 survey respondent, $\boldsymbol{\beta}$ is the column vector of regression parameters, and $\boldsymbol{x}$ is the vector of independent variables.

3.3 Modelling the Response Probabilities

The dataset for the 1986/87 panel of LMAS consisted of 66,817 individuals, of which 3,385 (5%) were nonrespondents to the 1987 interview. Demographic variables that were likely to be related to nonresponse were chosen from the 1986 LMAS master file as possible independent variables for the model.

The variables examined for inclusion in the nonresponse model were:

Province at 1986 interview
Urban/Rural area indicator at 1986 interview

Household size at 1986 interview
Type of dwelling (house; other) at 1986 interview
Status of dwelling (owned; rented) at 1986 interview

Sex
Age at 1986 interview
Marital status at 1986 interview
School attendance (full time; part time; none) in 1986
Highest level of education at 1986 interview

Any employment in 1986
Any unemployment in 1986
Any out-of-labour-force in 1986
Number of jobs in 1986
Any short tenure jobs (< 2 years) held in 1986
Any long tenure jobs (2 years or more) held in 1986
Any absences from work in 1986
Industry of job(s) in 1986

Average weekly income (over all jobs) in 1986
Received any unemployment insurance in 1986
Received any welfare in 1986
Moved (changed address between 1986 interview and 1987 interview)

All the categorical variables were converted to groups of dichotomous variables. The differences between respondents and nonrespondents with respect to the independent variables were analyzed. The correlations between all pairs of these variables were examined to find any potential multicollinearity.

First, a stepwise linear regression procedure was used to identify potentially useful variables for the modelling. This reduction in the choice of variables resulted in fewer variables to be entered into the logistic procedures saving considerable computer resources. The variables given in the STEPWISE procedure were entered into the SAS procedure PROC LOGISTIC with the BACKWARD and FAST options. These options allowed LOGISTIC to use an approximate backward elimination method to eliminate nonsignificant variables. Different logistic regression models were fitted to the full dataset using combinations of the most significant variables identified from the sample file. A consideration in choosing the model was the number of variables. It was desired to have a model with a small number of variables so that utilizing the model would be simple.

The model is used to make adjustments to the weights of the respondents in the second year (1987). For this model, the dependent variable was total nonresponse, and the independent variables were characteristics observed the previous year (1986) plus the current year's information (1987) on whether or not the person moved.

The BACKWARD option of PROC LOGISTIC was used with the sample file to identify eight variables related to nonresponse.

Male	(MALE)
Single	(SINGLE)
Rented dwelling	(RENT)

Any employment	(ANYEMP)
Highest education=secondary	(EDUCSEC)
Moved since 1986 interview	(MOVED)
Household size, to a maximum of 8	(HHS)
Age	(AGE)

Before fitting the models on the full dataset, the two continuous variables (household size and age) were examined for linearity on the logit scale. As with the prediction model, the age variable was replaced by two dichotomous variables for age: AGE1 for persons aged 25-54, AGE2 for persons aged 55-69 - the survey was conducted for persons aged 16-69 - and a transformation was applied to household size (HHSTRANS=|HHS-4.5|).

Four models were fitted to the full dataset: (1) using all eight variables; (2) using all except RENT; (3) using all except EDUCSEC; (4) using all except EDUCSEC and AGE. Although all eight variables were significant using the sample file, when the models were fitted to the full data file, certain ones no longer appeared important. However, it was decided to retain them in the models anyway. The statistics for evaluating the fit of the models indicated few differences between the four models. The Pearson residuals were plotted against the fitted values and the residual plots were examined. The residuals from Model (3) indicated a slightly better fit with fewer extreme values. Again using the sample file, the data were examined for the presence of two-way interactions between the variables in the model. Two sets of interactions were added to the model: the (AGE1 AGE2)*HHSTRANS and (AGE1 AGE2)*SINGLE. A summary of the fitted values for this model is given below. Note that the age and single variables as well as their interactions are not statistically significant. Nevertheless, when a model was fitted with these variables removed, it was found that there were more extreme values in the residuals.

Table 1

Parameter Estimates for Weighting Final Model.

Variable	$\hat{\beta}$	s.e.	χ^2
INTERCEPT	-3.81	0.14	702.59
HHSTRANS	0.13	0.06	4.97
MALE	0.25	0.04	41.98
RENT	0.23	0.04	29.14
SINGLE	0.11	0.16	0.43
MOVED	2.31	0.04	3065.95
AGE1	-0.15	0.17	0.75
AGE2	-0.19	0.15	1.65
AGE1*HHSTRANS	0.02	0.07	0.07
AGE2*HHSTRANS	0.05	0.06	0.55
AGE1*SINGLE	0.13	0.18	0.52
AGE2*SINGLE	0.11	0.17	0.40

Using the estimated parameters from the final model, predicted probabilities of nonresponse were calculated for all respondents to the 1987 interview and a nonresponse adjustment was made. Finally, a poststratification adjustment to population control totals at the province-sex-agegroup level, yielded the 1987 final weight.

3.4 Evaluation of the Weights

If the nonresponse weighting adjustment is adequate, there should be no difference in estimates obtained from the 1986 respondents and estimates obtained from the 1987 respondents when tabulating on 1986 characteristics. A number of demographic and labour-related characteristics were evaluated. Estimates were calculated using the 1986 weights, the 1987 model-adjusted weights, and the 1987 regular weights, including a ratio-adjustment at low geographic levels for nonresponse adjustment. For each characteristic a 95% confidence interval was calculated for the estimate based on the 1986 weights. The two 1987 estimates were compared for differences to the 1986 estimates as well as differences to each other. Tables 2 and 3 below show some of the results. Table 3 incorporates the poststratification adjustment, which, in general, improves the estimates.

Table 2

Comparison of the estimates with the two non-response adjustments, before the post-stratification, tabulated on 1986 characteristics.

	1986 estimate	95% c.i. for 1986 estimate	1987 model-based estimate	1987 regular estimate
Marital Status Married	64.6%	(64.1,65.1)	65.1%	65.7%
Single	26.7%	(26.3,27.0)	26.3%	25.7%
Widowed	3.1%	(2.9,3.3)	3.0%	3.0%
Divorced	5.7%	(5.4,6.0)	5.6%	5.5%
Highest Education Grade 0-8	14.7%	(14.2,15.2)	14.6%	14.6%
Secondary	50.3%	(49.7,50.9)	50.0%	50.0%
Some Post-Secondary	10.1%	(9.8,10.4)	10.2%	10.1%
Post-Sec. Cert./Dip.	12.9%	(12.5,13.3)	13.1%	13.1%
University Degree	12.0%	(11.6,12.4)	12.2%	12.2%
Weeks Employed in 1986 0 weeks	22.8%	(22.4,23.2)	22.6%	22.5%
1-26 weeks	12.0%	(11.7,12.3)	11.7%	11.6%
27-48 weeks	12.2%	(11.9,12.5)	12.1%	12.0%
49-52 weeks	53.0%	(52.4,53.6)	53.6%	54.0%

Of all the characteristics compared, only one 1987 estimate was outside the 1986 confidence interval: weeks employed=49-52 using the regular weighting. One pattern was clear, however. The estimates using the model-based weights were consistently closer to the 1986 estimates than those using the regular method of weighting.

4. FARM FINANCIAL SURVEY

4.1 Introduction

The Farm Financial Survey (FFS) has been a regular agricultural survey since 1980. The objective of the survey is to gather financial information on Canadian farmers. The survey collects information on revenues, expenses, assets and liabilities. Crop and livestock information are also collected to measure physical characteristics of the farms. Due to the collection of sensitive data, a low response rate has always been observed for the survey. A study was initiated on the 1992 survey data to identify the causes of nonresponse and possible solutions to reduce its impacts on the estimates.

The population of interest consists of all Canadian farms active for the reference year, excluding the multi-holding companies, the institutional farms, the community pastures, the farms on Indian Reserves and the farms with less than $2,000 in sales. The survey population is represented by a list frame and an area frame. The 1992 list frame was a register of all of the 1986 Census farms without the farms defined by the above exclusion rules. The list frame was stratified within each province by farm type and by farm size. The farm size was defined by the total farm assets derived on the Census.

The area frame was used to compensate for the undercoverage due to the Census itself or caused by new farms which started their activities since 1986. Basically, the area frame was a list of land segments outlined on topographic maps. Stratified replicates of segments were selected from the area frame. All farmers operating some land in the sampled segments were enumerated, and a register was created. There were 1,153 area frame farms that did not appear on the list frame. They were all contacted for the FFS as for other agricultural surveys. In addition to the area frame farms, a stratified sample was selected from the list frame to obtain a overall sample of about 12,000 farms. See Britney and Poirier (1992) for more details on the 1992 FFS sample design.

Domain estimation within each stratum was performed to obtain estimates of level from both the list and area samples. The simple expansion estimator was used on the 1992 list sample. The initial weighting was done by stratum using the population size over the observed sample size, so that a nonresponse adjustment is made at the stratum level. For the area frame, the estimation was done separately by replicate. For a given replicate, the data were aggregated at the segment level by applying to the farm data, factors corresponding to the proportion of the farms within the segment. Then, the segment totals received expansion weights (π) to represent the population. When nonresponse occurred for an area farm, the respondents within the same segment were reweighted on an area basis to compensate the farm land for which data were unavailable. For both the list and area units, partial nonresponses were donor imputed and used the same way the regular respondent were. Details are given by Maranda (1989).

The nonresponse observed in the 1992 Farm Financial Survey was relatively important. The FFS questionnaire was relatively long

with many sensitive questions related to the financial balance sheet. The resulting total unit-level refusal rate of about 15% across the country was the highest of our agricultural surveys. In addition to the total refusals, the no-contacts represented another 5% of the sample. Some provinces presented higher nonresponse rate than others. In Saskatchewan, data were unavailable for almost 30% of the sampled farms. Table 4 shows the nonresponse distribution across the country.

Table 4

1992 Nonresponse Distribution

Province	Sample Size	Total Refusal	No-Contact
Newfoundland	211	20	16
P.E.I.	528	64	14
Nova Scotia	668	74	11
New Brunswick	537	48	18
Quebec	1311	124	51
Ontario	1513	250	84
Manitoba	1756	321	109
Saskatchewan	1880	424	126
Alberta	1868	312	109
B.C.	1448	175	138
Total	11720	1812	676

4.2 Nonresponse Models

A part of our study was first to identify the causes of nonresponse. This could help taking decisions related to the collection methods to increase the response rate. It also allowed the identification of factors that may be considered in any nonresponse reweighting models. Since, the no-contacts and the refusals were possibly caused by different factors, they were kept separate in all of the hypotheses we made. The potential causes that were studied on the 1992 FFS data are:

1 - **The frame origin:** This corresponded to whether or not the farm was selected from the list frame. Since the area frame farms were conceptually missed by the Census, they probably showed characteristics that trend to generate nonresponse. Also, since the area sample is being used by many agricultural surveys, the area frame farms might refuse because of their response burden.

2 - **The farm size:** The capability or the will to respond could depend on the farm organisation and on its size.

The size was evaluated using the farm assets and sales obtained from the 1986 Census of Agriculture. This size was available only for the list units.

3 - **Geography:** The geographic location aimed to identify the interviewer effect and the impact of farmer associations which could boycott government surveys because they were not benefited by their programs. Census divisions were used to verify this hypothesis.

4 - **Farm type:** The farmer's availability depends on the type of his farm. Seven categories of farm type were used to differentiate the farms.

5 - **Response burden:** Because the large number of agricultural surveys held in a short period of time, the response burden became important for some farmers. The overlaps with the December Stock Survey and the January Livestock Survey (JLS) were both studied to verify its impact on the response rates. These surveys were both conducted less than two months before the FFS. The effect of the overlap with the previous FFS, held in 1990, was also investigated.

6 - **Age of operator:** The age of an operator could affect its will to cooperate, but the data available to verify this hypotheses were not reliable enough to do any studies.

Tests of independence were conducted to verify if any of the above factors could affect the response status: 'completed', 'no-contact' and 'refusal'. The partial refusals were included with the completed questionnaires. The statistic used to conduct the independence tests was the weighted Pearson statistic χ^2 with the Fellegi (1980) correction to take into account the design effect. This test is known to be conservative.

The farm assets and sales, which were both indicators of the farm size, were replaced by categorical variables defined using the estimated quartiles. The census divisions representing the geographic location were grouped into a maximum of 9 classes within each province. This ensured a minimum number of observations within each cell of the cross classification with the response status.

In some cases, where dependence was detected between the factors and the response status, additional tests were conducted to identify the nature of the dependence. This was done through statistical tests on proportions. The most important conclusions are described here but more details can be found in Poirier (1994).

The independence tests, conducted with a confidence level of 5%, identified certain causes of nonresponse. First, within each province except Ontario, the farm type had a high impact on nonresponse. Also, the farm size measured in term of assets affected the response rates in most of the provinces, but no significant impact was due to the sales variable. The geographic location and the response burden generated by the previous FFS survey significatively affected the probability to respond in three provinces. Finally, the frame origin and the overlap with the January Livestock Survey or the December Crops Survey seemed to not affect the response status at all.

As in Section 3, we modelled the nonrespondent behaviour with logistic regression modelling using the SAS procedure LOGISTIC. We performed the analysis separately by province. Using frame origin as an independent variable, the results confirmed the previous conclusions of no frame effect. Since some variables were not available for the area sample and since the frame origin did not seem to affect the response, the remaining analyses were performed only on the list units, which represented more than 90% of the whole sample. In the rest on this paper, the results applied for the list units only.

For the purposes of this study, the following variables were included in the model:

i) Assets (1 if assets are smaller than the median, 0 otherwise),
ii) Sales (1 if sales are smaller than the median, 0 otherwise),
iii) $Type_i$ (1 if in the i^{th} farm type, 0 otherwise),
iv) $Area_i$ (1 if in the i^{th} geographic area, 0 otherwise),
v) FFS (1 if in the 1990 FFS sample, 0 otherwise),
vi) JLS (1 if in the 1992 JLS sample, 0 otherwise),

The farm types are (1) crop farms, (2) dairy farms, (3) cattle farms, (4) hog farms, (5) poultry farms, (6) sheep farms, and (9) unknown type of farm.

The variables that were found more significant by the BACKWARD option within the provinces were kept in the model. The most commonly selected variables were the farm types and the FFS variables. Table 5 shows the resulting estimated parameters corresponding to the variables kept in the model. It also provides the attached χ^2 values.

Table 5

Nonresponse Logistic Parameters with their χ^2 Values

Prov.		Intercept	FFS	Assets	Type4	Type5	Type6	Type7	Area1	Area2	Area3	Area5	Area6	Area7
NFLD	β	1.37	1.05	.	.	-1.78	.	.	.	.	.	.	.	.
	χ^2	37.23	5.75			6.76								
PEI	β	1.75	.	.	.	.	.	.	.	.	.	.	.	.
	χ^2	204.18												
NS	β	1.93	.	.	.	.	.	.	.	.	.	.	.	.
	χ^2	275.05												
NB	β	1.65	0.71	.	.	.	.	.	.	.	.	.	.	.
	χ^2	98.87	6.79											
QUE	β	2.05	.	.	.	-0.58	.	.	.	-0.45	.	.	.	.
	χ^2	392.98				5.11				4.03				
ONT	β	0.93	0.53	.	.	.	1.07	.	.	.	.	.	.	.
	χ^2	71.00	14.18				4.09							
MAN	β	0.94	0.44	.	.	-0.45	.	-1.16	0.59	.	.	.	.	.
	χ^2	113.35	13.34			4.21		38.90	14.05					
SASK	β	0.94	0.51	0.31	.	.	.	-2.40	.	.	.	.	-0.61	-0.32
	χ^2	78.44	18.67	6.85				134.31					12.20	4.22
ALB	β	1.04	0.69	0.26	-0.33	.	.	-1.12	.	.	.	.	.	.
	χ^2	79.25	30.12	4.30	5.51			40.48						
BC	β	0.88	.	.	.	.	.	.	0.93	0.86	0.47	0.43	.	.
	χ^2	74.82							15.29	21.01	5.85	5.13		

Variables that were not significant for any of the provinces have been removed from this table. From the χ^2 results, it appears that the FFS overlap and the farm Type7 (representing the sheep farms) have the most important impacts on nonresponse. The positive FFS parameters mean that farms overlapping the previous FFS tended to have higher response rates, whereas the negative sheep farm parameters (Type7) imply they tended to respond less often.

Weighted regressions were also fitted to the data using the WEIGHT statement of the LOGISTIC procedure. The weighting variable was defined at the stratum level as the design weight adjusted to the overall sample size. Stratum level adjustments were not performed. The resulting estimated parameters were very close to the first set of estimates which, as we explained in Section 3, is highly desirable.

4.3 Evaluation of the Weights

To evaluate the nonresponse adjustment, the 1992 frame values representing farm assets were estimated from the sample. Assets levels were estimated for each province with the corresponding

coefficient of variation (CV), including the nonresponding units. Then, estimates based only on respondents were produced, using the original weight, adjusted for nonresponse at the stratum level only. By comparing both set of estimates we could derive the nonresponse bias introduced by the current method. Finally, regression adjusted estimates were produced from the above logistic model.

If we denote by $\hat{Y}_0$ and CV_0 estimated level and coefficient of variation, respectively, from the full sample, and $\hat{Y}_{adj}$ the corresponding adjusted estimates based only on respondents, we can estimate the bias associated with the adjustment model. In Table 6 we show the results.

Table 6

Comparison of the Adjustment Models for 1992 Frame Value of Farm Assets

Prov.	Y_0	CV_0 (%)	Stratum Adjusted Weight	Logistic Adjusted Weight
			BIAS (%)	BIAS (%)
NFLD	7.7 E07	2.91	2.34	1.80
PEI	6.7 E08	0.76	-0.35	0.19
NS	7.9 E08	0.68	-1.15	-0.27
NB	6.1 E08	0.82	-0.16	0.46
QUE	8.5 E09	0.53	-0.28	-0.04
ONT	2.1 E10	0.56	-1.16	-0.85
MAN	9.0 E09	0.57	0.21	0.45
SASK	2.7 E10	0.52	-0.40	-0.89
ALB	2.7 E10	0.57	0.53	-0.12
BC	5.6 E09	0.62	-2.32	-2.24
TOTAL	1.0 E11	0.25	-0.35	-0.54

We see that the logistic adjusted weight generally performs better, but not consistently so. In fact the bias increases for NB, MAN, SASK, and the Total. To improve the model, inclusion of some interaction factors like size and farm type, or size and geography was tried but they were rarely kept in the model and when they were, the resulting effects were small and their impact was negligible.

4.4 Conclusion

The selected model did not consistently provide the expected bias adjustment. This may be caused by a low number of factors included in the model or by the fact that significant factors were used in the frame stratification. Future work might include looking for more interactions using the Automated Interaction Detection method used in Section 3. Also, now that the 1993 data are available, the study could be extended.

5. SUMMARY

Nonresponse adjustment through reweighting is now in common use. We have shown that the success of this technique generally depends on having available variables that can be used as good predictors of the nonresponse behaviour. Having such variables, various models can be used to adjust the estimates based on the predicted response propensities. This seems to be the best general approach. Other approaches include using estimation methods such as regression estimators to compensate for the deficiencies of the sample. We have seen that if the regression models are valid, the nonresponse bias vanishes.

We have concentrated here on asymptotic biases. However, there are still many unresolved issues for estimation of variances and construction of confidence intervals. As well, we have not properly addressed the issue of whether or not to use the sampling weights when fitting the nonresponse models. In our examples, the weighted and unweighted versions of the estimated response models gave similar results. This is highly desirable since it confirms the validity of the model.

Nonresponse problems will not go away. A better understanding of the response mechanisms will lead to better survey practices in the long run.

ACKNOWLEDGEMENTS

We are especially grateful to Graham Kalton, who most kindly provided us with a comprehensive list of references on this topic. We would also like to thank Ralph Folsom and Steven B. Cohen for sending us some of their recent work.

REFERENCES

Alexander, C.H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology* 13, 183-198.

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics* 4, 251-260.

Binder, D.A. and Théberge, A. (1988). Estimating the variance of raking-ratio estimators. *Canadian Journal of Statistics* 16, 47-55.

Britney, H. and Poirier, C., (1992). 1992 Farm Credit Corporation: Design documentation, Internal Paper, Agriculture Section, Business Survey Methods Division, Statistics Canada.

Chapman, D.W., Bailey, L., and Kasprzyk, D. (1986). Nonresponse adjustment procedures at the U.S. Bureau of the Census. *Survey Methodology* 12, 161-180.

Deville, J-C., Särndal, C-E. and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association* 88, 1013-1020.

Ekholm, A. and Laaksonen, S. (1991). Weighting via response modeling in the Finnish Household Budget Survey. *Journal of Official Statistics* 7, 325-337.

Eltinge, J.L. and Yansaneh, I.S. (1993). *Weighting adjustments for income nonresponse in the U.S. Consumer Expenditure Survey*. Technical Report No. 202, Department of Statistics, Texas A&M University, College Station, Texas.

Fellegi, I.P. (1980). On adjusting the Pearson chi-square statistic for cluster sampling. *Journal of the American Statistical Association* 71, 665-670.

Folsom, R.E. (1991). Exponential and logistic weight adjustments for sampling and nonresponse error reduction. *Proceedings of the Social Statistic Section*, American Statistical Association, 197-202.

Iannacchione, V.G., Milne, J.G. and Folsom, R.E. (1991). Response probability weight adjustments using logistic regression. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 637-642.

Judkins, D. and Lo, A. (1993). Components of variance and nonresponse adjustment for the Medicare Current Beneficiary Survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Lepkowski, J., Graham, G., and Kasprzyk D. (1989). Weighting adjustments for partial nonresponse in the 1984 SIPP Panel. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 296-301.

Maranda, F. (1989). Proposal for NFS estimation, Internal Paper, Agriculture Section, Business Survey Methods Division, Statistics Canada.

Oh, H.L. and Scheuren, F. (1983). Weighting adjustments for unit nonresponse. In *Incomplete Data in Sample Surveys, Volume 2: Theory and Bibliographies* (W.G. Madow, I. Olkin and D. Rubin, eds.), 143-184. Academic Press, New York.

Poirier, C. (1994). The causes of non-response in the context of agricultural surveys. Internal Paper, Agriculture Section, Business Survey Methods Division, Statistics Canada.

Särndal, C-E. and Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review* 55, 279-294.

Särndal, C-E., Swensson, and Wretman, J. (1992), *Model Assisted Survey Sampling*. Springer-Verlag, New-York.

Singh, M.P., Drew, J.D., Gambino, J.G., and Mayda, F. (1990). *Methodology of the Canadian Labour Force Survey 1984-1990*. Statistics Canada, Catalogue 71-256, Ottawa.

NONRESPONSE DISCUSSION

J. Michael Brick
Westat, Inc.

I would like to thank the authors of both papers for presenting insights on very different aspects of the nonresponse problem. The first paper, "Exploring Nonresponse in U.S. Federal Surveys" by Gonzalez, Kasprzyk, and Scheuren, reports on the activities of a Subcommittee on Nonresponse for the Federal Committee on Statistical Methodology (FCSM). The main focus of this paper is on trends in response rates for federal surveys over the last 10 years. The second paper, "Model-based Reweighting for Nonresponse Adjustment" by Binder, Michaud and Poirier, reports on research in a particular approach to reduce the bias due to nonresponse.

In "Exploring Nonresponse in U.S. Federal Surveys," the authors describe recent investigations of the levels of response rates and trends in nonresponse and refusal rates in federal surveys conducted from 1982-1991. The other parts of the mission of the subcommittee include reporting on correlates of response rates and other related matters. We look forward to reports on those activities later this summer.

The findings on trends in nonresponse rates are especially welcome. In general, these data show that there has been no increase in nonresponse or refusal rates over the last decade. This finding contradicts the conventional wisdom that nonresponse rates, and refusal rates in particular, have been rising due to the increases in the burden on respondents. It clearly shows that it is risky to make general conclusions about this type of phenomenon based on limited data.

Given that response rates seem to be fairly stable over the last decade, the question still remains about the value of judging the quality of survey data over time from this type of data. After all, we are really interested in the quality of the data, not the response rates themselves. However, response rates are the only measure of quality typically produced from surveys. While having something that is measurable is useful, response rates are frequently not very informative of the quality of the data.

Using what Deming calls the "modern" approach to quality control, it is important to recognize that the response rates are really a <u>product</u> characteristic of a survey. Survey methodologists, on the other hand, must concentrate on the <u>process</u> of collecting data. The danger is that examining product variables alone can mask the process related to the quality of the data. For example, if the population surveyed changes so that a group that has generally high response rates is now sampled, then we might expect the response rates to increase. Thus, response rates might change even if respondents willingness to participate remained the same. Similarly, unweighted response rates may be affected by changes in the sample design, such as changes in the sampling rates for different segments of the population.

For survey methodologists, the process must be the most important part of their job. They are responsible for producing survey data of the highest possible quality within the constraints of the survey. In many ways, this paper only peripherally addresses this set of core users, because it does not discuss the process itself. To do this requires more process data, much of which is not comparable across surveys.

The focus on response rates is also limited for other reasons. The important relationship between response rates and costs is not explored in this paper. Cost data are crucial, since the cost of obtaining the same level of response rates may have increased or decreased. Unfortunately, cost data are extremely difficult to obtain and formulate in a manner that is useful for comparison

purposes. Developing a general method for analyzing cost data remains, in my opinion, one of the most important unresolved problems in survey research.

As the authors of this paper indicate, improving response rates does not always improve the quality of the survey estimates. This has been misinterpreted by some researchers to mean that response rates are not important. They are important and reasonable efforts should be made to eliminate nonresponse. The real question is related to how much effort should be placed on improving response rates versus reducing other sources of error in a survey. The answer to this question is not always clear, but there are guidelines for reasonable practice. In general, the resources devoted to a source of error (be it sampling or nonsampling) should be proportionate to the size of that error relative to the sum of sampling and nonsampling errors.

In one of their recommendations, Gonzalez, Kasprzyk, and Scheuren state that research should be encouraged by survey administrators as a way to improve the effectiveness of data collection operations. They include a specific recommendation that studies of collecting items to improve post-survey adjustments should be encouraged. I think this is a very interesting choice for the recommendation. It supports collecting items rather than exploring new methodologies. I applaud this direction. Research on collecting additional items is a sound direction to improve our ability to decrease nonresponse bias.

The paper by Binder, Michaud and Poirier, "Model-based Reweighting for Nonresponse Adjustment" is very consistent with the recommendations of the FCSM Subcommittee. Since the type of nonresponse discussed in this paper is not planned, randomization does not apply and a modeling approach must be used to reduce the bias. The authors discuss the results from applying a specific form of modeling nonresponse for two different surveys.

Their efforts to reduce the bias in the estimates for the Survey of Labour and Income Dynamics (SLID), closely parallel work done on the U.S. Survey of Income and Program Participation (SIPP). The Census Bureau sponsored several efforts to examine methods to reduce attrition bias in the SIPP and many of the papers from these efforts will be presented this summer. In the SIPP research conducted at Westat, we used many of the same techniques to model the nonresponse as Binder, Michaud and Poirier, with much the same result.

I would like to make two specific comments on the paper prior to returning to a discussion of the importance of collecting items for reducing nonresponse bias. The first deals with the use of the logistic regression model used to predict response propensity. This work, and our own efforts in the SIPP research, suggest that these types of models are not the most conducive to uncovering relationships between variables when faced with a large number of potential predictor variables. Categorical search algorithms seem better equipped at identifying nested relationships that might be important for reducing nonresponse bias. In this paper, these relationship were uncovered with the search algorithms and added to the logistic regression model. We have used the output from the search algorithms to form cells for nonresponse adjustment directly. Both methods seem to work reasonably well.

The second comment is really a question. The item most related to nonresponse was mover status. People who moved were far more likely to be nonrespondents. This characteristic has obvious appeal and could be of great importance in nonresponse bias reduction. The question is how nonrespondents were defined as movers or nonmovers based on the data available. In most surveys, we do not have these data for nonrespondents. I think it would be useful for other survey methodologists to see how the authors made this classification for SLID.

If an item such as whether or not the person has moved is available, then it can be used to estimate response propensity. Having items that are good estimators of response propensity is critical for reducing bias. I have found that different nonresponse adjustment methods, when

properly applied, often have approximately the same performance. This may happen because almost all the methods estimate individual response propensities conditional on auxiliary data. Estimating a response propensity for each unit is very difficult since only one observation is available (the unit either responded or did not respond). Items related to the likelihood of responding at the individual level are, therefore, extremely valuable to improve this estimation.

The authors take an approach that seems most sensible. They include as large a number of items as possible in the model without allowing the adjustments to become so variable as to unduly increase the variance of the estimates. Since nonresponse bias is a function of the covariance between the estimate and the response propensity, items related to both should be entertained. As suggested in this research, the first items included should be those related to response propensity. If it were possible to model these response propensities precisely, then the nonresponse bias for all the estimates would be eliminated. Since the modeling is imperfect, additional items highly related to key statistics from the survey should also be included, where possible.

While longitudinal surveys and studies in which there are frames with substantial data on the sampled units may be able to use this method to reduce the bias due to nonresponse, the more typical situation may be that faced in the Farm Financial Survey. In this case, the data on the frame are limited and do not appear to be predictive of either response propensity or the key estimates. The methods used are not very effective on reducing the nonresponse bias, as might be expected. If the nonresponse is large enough in a survey of this nature, the recommendation of the FCSM to develop a mechanism for collecting items useful for nonresponse adjustment should be seriously considered. This common situation reinforces the importance of keeping nonresponse to a minimum.

DISCUSSION

Joseph L. Schafer
Pennsylvania State University

1. Comments on "Exploring nonresponse in U.S. Federal Surveys"

This paper by Gonzalez, Kasprzyk, and Scheuren (GKS) is worthwhile reading. Section 2 gives a nice historical overview of the development of statistical methods for survey nonresponse. As this section clearly demonstrates, many of the great strides in the practice of nonresponse adjustment, and survey sampling in general, have come about as a direct result of personal interaction between statisticians in federal agencies and in those in academic circles. It is our hope that this type of fruitful interaction will continue in the years ahead.

In the world of surveys, there are many different types of nonresponse. Unit nonresponse arises when the entire vector of survey variables for a sample unit is missing. Item nonresponse arises when individual elements of the vector are missing. As pointed out in the other paper in this session, panel surveys often suffer from wave nonresponse, which occurs when the entire vector of survey variables is missing for a unit at a particular occasion or wave. The GKS paper has introduced a new type of missingness: nonresponse nonresponse, which occurs when members of the federal statistical community fail to report their nonresponse rates to members of the FCSM subcommittee. An even worse type of nonresponse nonresponse occurs when producers of federal surveys do not report basic information on nonresponse--such as nonresponse rates and methods of adjustment--to the users of their data.

It is worthwhile to ask whether nonresponse nonresponse is ignorable, in the sense defined by Rubin (1987). In the FCSM subcommittee's study, the nonresponse nonresponse would be ignorable if the nonresponding statisticians' surveys were representative of all federal surveys in terms of missingness rates, methods of adjustment, etc. Chances are, the nonresponding statisticians had desks that look like mine--paperwork stacked up a foot high all around--and they didn't find the time to return the questionnaire. Or, perhaps they didn't respond because they hadn't been tracking basic information such as nonresponse rates, and compiling the information would have required an unusual amount of effort. Whether those characteristics of the nonrespondents are systematically related to the basic study variables is anyone's guess. In the common, non-technical sense of the word, however, this nonresponse to the FCSM subcommittee's study is

probably not ignorable; we shouldn't ignore it. If members of the subcommittee found it so difficult to obtain even basic information on nonresponse from the very people who produce the survey, imagine how hard it must be for the average data user to do the same thing.

It was astonishing to see that almost half of the establishment surveys in the study had no tracking at all of the various components of nonresponse. We hope that in the future, producers of federal surveys will take to heart the recommendations of this report and devote a little more time and effort to studying and documenting the levels and causes of nonresponse. At any rate, it is comforting to learn that the basic perception that many of us had--that nonresponse rates have been rising over the past few years--may be only a perception, and the sum total of the various nonresponse components may not have changed very much.

Reading this paper made me think a lot about the future, and what directions we should take in our research on nonresponse and nonresponse adjustment. The suggestion near the end of the paper--that data collectors and statisticians work together as a team, sharing information in a way that is mutually beneficial--is especially thought-provoking. If this were done, it would open up entirely new avenues for developing improved methods of nonresponse adjustment. As shrinking budgets force us to re-allocate our resources, it may not be possible to continue to spend so much money chasing after nonrespondents, trying to get them to hand their data vectors over to us. It's becoming increasingly likely, for example, that the Census Bureau will not have the resources in the year 2000 to follow up on every housing unit that doesn't mail back its census form with a personal interview. Statisticians and data collectors should start to think long and hard about which nonrespondents they should attempt to follow up, and how persistently they should attempt to do so. They need to weigh the costs and benefits involved in converting refusals and not-at-homes to responses, so that they may decide whether the money might be better spent elsewhere.

In regard to this point, I would like to make two basic observations. The first observation is that not all nonrespondents are equally important. In industrial settings, statisticians have to carefully design their experiments. Trial runs are typically expensive, and they have to decide under what values of X they should collect their next value of Y. They know that certain, carefully chosen combinations of Xs will help them to "nail down" the quantities of interest much better than other combinations of Xs, so they choose their values of X carefully before spending additional money to collect another Y.

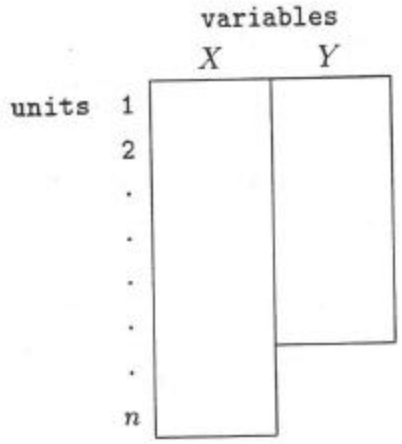


Figure 1: Data after the initial phase of collection

In a survey setting, there is typically an initial phase of data collection in which every unit in the sample has been contacted once, or at least an attempt has been made to contact them once. After this initial phase there will be a pool of nonrespondents, and there will be (a limited amount of) money to spend on attempts to convert some of them to respondents. Which nonrespondents should we go after first? To answer this question, we should look at their Xs. At this point, the data will resemble the diagram in Figure 1. Variables available from the sample frame, denoted by X, will known for all sample units. Survey variables, denoted by Y, will be available for the respondents only. Let us partition Y as $Y = (Y_{obs}, Y_{mis})$ where Y_{obs} denotes the values of the survey variables for the respondents, and Y_{mis} the values of the survey variables for the nonrespondents. It should be possible to build a regression model for the distribution of Y_{obs} given X, and then use this model to guess or predict what the values of Y_{mis} will be. These predictions may then form the basis for ranking the nonrespondents according to the priority with which they should be followed up.

Let $P(Y_{obs}|X,\theta)$ denote the form of the model fit to Y_{obs} given X, where θ represents some unknown parameters. Let y_i^* denote an element of Y_{mis}. To the classical survey statistician, y_i^* is a fixed, unknown constant. To the Bayesian, however, it is unknown and therefore a random variable. The uncertainty about y_i^* can be expressed Bayesianly as

$$V(y_i^*|X,Y_{obs}) = EV(y_i^*|X,Y_{obs},\theta) + VE(y_i^*|X,Y_{obs},\theta),$$

where the outer moments are taken with respect to $P(\theta|X,Y_{obs})$, the posterior distribution of the unknown model parameters. The first component on the right-hand side, $EV(y_i^*|X,Y_{obs},\theta)$, represents the residual variation of y_i^* about its predicted value from the regression model. The second component,

$VE(y_i^* | X, Y_{obs}, \theta)$, represents the uncertainty in the regression prediction itself. This decomposition suggests that if we want to minimize our uncertainty after a limited amount of followup, we should target as high priority those units that (i) have a high amount of residual variance, and (ii) have high leverage for estimation of θ. In other words, we should try to follow up the units (i) whose values of Y cannot be predicted well by our model, and (ii) whose values of X are unusual and thus, if they were converted to respondents, could greatly improve our ability to predict the missing Y values for the other nonrespondents.

In addition to the predictive variance of y_i^*, we also need to consider the probability that a nonresponding unit can successfully be converted to a respondent. Even if the predictive variance for a particular unit is high, it may not make sense to attempt followup if the followup operation is likely to be unsuccessful. This suggests construction of another regression model to predict the probability of successful followup--or, perhaps, the cost of successful followup in terms of number of attempts, field worker time, etc. Data for fitting this model might come from similar survey operations of the past, perhaps updated by data from the current survey as they become available.

As data collectors begin to share information with statisticians on an ongoing basis, one can imagine the development of a continuous-loop feedback system in which the field-operations unit provides data on respondents as they become available, and the statistical unit processes the information, updates the parameters of its regression models, and decides which of the remaining nonrespondents should be designated for followup.

The second general observation that I would like to make is that not all nonresponse mechanisms are the same. From a theoretical standpoint, it is useful to classify nonresponse mechanisms into two categories: mechanisms that are ignorable and mechanisms that are nonignorable. Using the notation developed above, an ignorable mechanism is one in which the probabilities of response do not depend on Y_{mis} after accounting for dependence on X and Y_{obs}. Ignorable nonresponse mechanisms tend to be easier to deal with than nonignorable ones, and virtually all methods of nonresponse adjustment in use today make some implicit assumptions of ignorability.

From a practical standpoint, however, nonresponse mechanisms should probably be classified into a slightly different dichotomy: mechanisms that are known to be ignorable, versus mechanisms that are not known to be ignorable. Mechanisms that are known to be ignorable include those in which the missing data are missing by design. Surveys that employ double sampling,

matrix sampling, etc. result in rectangular datasets with patches of missing data that are missing by design; the data are unrecorded because the data collector never intended to collect them. When data are missing by design, ignorable missing-data techniques may be applied without fear of introducing bias. The more insidious type of missingness mechanism is the unknown type. When the nonrespondents are a self-selecting subsample, we do not really know how strongly the selection process may be related to the missing data Y_{mis}. When faced with missingness of this type, the only thing that a practitioner can usually do is to apply some ignorable missing-data technique and hope for the best--i.e. pray that any biases incurred by nonignorability will not be severe.

As long as resources for data collection are finite, a certain amount of missing data will be inevitable. But the point that I want to make is this: By intelligent allocation of resources in the followup operation, we may be able to convert a substantial amount of the missing data that would ordinarily be of the type "unknown" to the type "ignorable." In a typical followup operation of today, attempts are made to follow up nonrespondents in a rather haphazard (i.e. unplanned) fashion until the resources run out, at which time the data collectors close out their operation and get on with their lives. Decisions about which nonrespondents are to be followed up are not made by a central decision-making unit, but are made in the field by supervisors or by the interviewers themselves. It may be that the field staff is placing high priority on the nonrespondent units that appear to be easy to get, thereby attempting to minimize the number of nonrespondents that remain after closeout. Although minimizing the nonresponse rate is a laudable goal, the end result is that all of the nonresponse that remains after closeout is of the type "unknown." From a statistician's point of view, a better strategy may be to concentrate one's resources on obtaining data for a probability sample of nonrespondents, a sample that is guaranteed to be representative of the nonrespondent pool. Even if data for these units are expensive to obtain--e.g. requiring a large number of call-backs--and the overall rate of missingness in the end is higher than it would have been if the followup decisions were made by field staff, the end result will be that the missing data for nonrespondents that are not included in the followup sample will be of the type "ignorable."

As a scientist, I would be willing to trade a few percentage points of missingness for a guarantee that (at least most of) the missing data are ignorable. I suspect others would as well. The tradeoff between the cost of missing data versus the benefit of knowing the missingness mechanism is a subtle but important issue to which statisticians ought to pay more attention

in the future.

2. Comments on "Model-based reweighting for nonresponse adjustment"

This paper by Binder, Michaud, and Poirier (BMP) discusses in detail the methods used by StatCanada to model response propensities in two of its ongoing surveys. By reading this paper, and examining the tables of regression coefficients, one develops an excellent sense of what factors may be related to nonresponse in demographic and establishment surveys. It is interesting to note that in regard to these two surveys, BMP and StatCanada seem to be following the recommendations of the authors of the previous paper: clearly documenting the rates of nonresponse, the factors related to nonresponse, and the methods used for nonresponse adjustment.

Throughout this paper, the value of adopting a model-based approach to nonresponse adjustment clearly shines through. By constructing an intelligent model for the nonresponse mechanism, one is able to carry out a nonresponse adjustment using many more explanatory variables than would otherwise be possible using a more traditional approach. In a more traditional approach, one would form adjustment cells by crossing the classes of every explanatory variable. This would be equivalent to building a response-propensity model that includes all possible interactions among the explanatory variables, whether or not those interactions have much predictive power (and they often don't). The modeling approach adopted by BMP allows them to exclude unimportant high-order interactions, and instead include main effects for a larger number of explanatory variables.

One issue that may deserve a little more attention is what should be done with the estimated response propensities once they are calculated. On this point, statisticians north of the U.S.-Canada border tend to use the reciprocals of these probabilities as factors in the nonresponse weighting adjustment. Statisticians south of the border tend to form classes--e.g. by dividing the estimated propensities into quintiles--and reweight the observations within these classes. Little and Rubin (1987) comment that the latter may sacrifice a little bias for the sake of reduced variance and robustness against model failure. I wonder if anyone has done a comparison of the two methods in a realistic setting to see which one tends to perform better.

Another important issue, which is perhaps beyond the scope of the BMP paper, relates to the underlying philosophy of response-propensity weighting. Response-propensity weighting attempts to control and reduce nonresponse bias. The theory of propensity scores (Rosenbaum and Rubin,

1983) says that if reweighting could be performed on the basis of the actual (as opposed to estimated) propensity scores, then the reweighted distribution of the respondents would not be systematically any different from that of the respondents--in other words, nonresponse bias would be eliminated. But of course, bias is only one component of error, the the other being variance. As pointed out by Little (1986), response-propensity weighting may do very little to control variance. One might also want to consider forming weighting classes on the basis of variables that are highly correlated with the survey variable of interest (if any are available), so that variance might also be reduced. Perhaps forming weighting classes on the basis of two variables--a linear predictor of the response propensity, and a linear predictor of the survey variable of interest--may be a reasonable approach.

The methods described in the BMP paper also raise a number of theoretical issues that deserve a closer look by statisticians in the future. One question involves the use of complex, automated variable-selection procedures to choose a model for nonresponse adjustment. Whenever the form of the model is chosen through examination of the sample data, the procedure used to select the model should be considered a part of the overall method of nonresponse adjustment. Any calculations of bias and variance--whether carried out analytically, or by simulation, jackknifing, etc.--should recognize that the model itself is sample-dependent and therefore random, and the model selection procedure must therefore be included in the calculation.

Another, perhaps more basic, issue pertains to the criteria used for selecting the model. BMP emphasized the principle of parsimony, eliminating variables whose effects were not significantly different from zero. They also included variables tended to reduce the number of extreme residuals. In the end, they were left with models that had very few parameters relative to the size of the dataset. I was left with the feeling that they could have included more variables, provided that the model-fitting could be accomplished in a reasonable amount of time. The usual criteria given by textbooks on regression modeling--high R^2, low prediction error, all coefficients statistically significant, and so on--are usually appropriate when the goal is to acquire some scientific understanding of how the response variable is related to the pool of potential predictors. When the goal is not necessarily scientific understanding, but adjusting for nonresponse, however, it is not yet clear what model-selection criteria statisticians ought to be using.

Finally, another issue that deserves further investigation is the proper

role of sample-design information in the construction of response-propensity models. I suspect that many statisticians would attempt to include design information by fitting logistic regressions with standard software, including the case weights (inverses of the sample-selection probabilities) in the fitting procedure. The correctness of such a procedure is not at all clear. If the goal were to estimate regression coefficients for predicting nonresponse for the entire population, then including the case weights would be appropriate. The goal in response-propensity modeling, however, is to estimate the probability of response for the units in the current sample. To the extent that this response propensity is related to covariates describing the sample design (e.g. stratum or cluster indicators), those covariates ought to be included in the model somehow. But merely weighting the cases by their inverse probability of selection is probably not sufficient to guarantee that the special features of a dataset that arise from complex sampling, such as clustering effects, are appropriately described.

Additional references

Little, R.J.A. (1986), "Survey nonresponse adjustments for estimates of means," International Statistical Review, 54, 139-157.

Rosenbaum, P.R. and Rubin, D.B. (1983), "The central role of the propensity score in observational studies for causal effects," Biometrika, 70, 41-55.